DOCUMENT RESUME

ED 173 400 TM 009 442

AUTHOR Reckase, Mark D.
TITLE Item Pool Construction for Use With Latent Trait Models.
PUB DATE Apr 79
NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; *Item Banks; *Mathematical Models; *Sampling; Senior High Schools; *Test Construction; Testing Problems
IDENTIFIERS *Latent Trait Models; One Parameter Model; Sample Size; Test Length; *Test Linking; Three Parameter Model

ABSTRACT

Because latent trait models require that large numbers of items be calibrated or that testing of the same large group be repeated, item parameter estimates are often obtained by administering separate tests to different groups and "linking" the results to construct an adequate item pool. Four issues were studied, based upon the analysis of a realistic application of several linking procedures: (1) required sample size; (2) number of items overlapping between linked tests to allow accurate translation of scale; (3) effect of linking on quality of existing item parameter estimates; and (4) best linking procedure for the two latent trait models studied (one-parameter and three-parameter logistic models). The linking techniques used were the major axis method; least squares; least squares with outliers deleted; and maximum likelihood. Three different calibration programs were used. Data were achievement test scores for 4,000 high school students on the 357-item Iowa Tests of Educational Development (ITED). Results of the analyses indicated that no method was clearly superior. The more simple one-parameter logistic model was found to be more advantageous as far as required sample size. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Mark D. Reckase

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED173400

TM009 442

Item pool construction for use with latent trait models

by

Mark D. Reckase
University of Missouri - Columbia

The development of latent trait models to describe the interaction of a person with a test item has resulted in many new opportunities for aptitude and achievement testing. The relatively sample free nature of the item parameters has made it convenient to produce algorithms that select and score items through the use of the computer. The recent developments in computer assisted test construction and tailored testing are examples of the application of this newly developed technology. However, if these new procedures can ever be expected to improve upon current testing methodology, large item banks must be developed so that the selection algorithms may be used to full advantage. Often item banks with an excess of 200 calibrated items are required (Reckase, 1976; Jensema, 1975). Unfortunately, it is seldom possible to administer the required number of test items to one large group so that the latent trait item parameters can be obtained through a single analysis. Moreover, the repeated testing of the same large group is usually inconvenient, if not impossible. Therefore, item parameter estimates are often obtained from the administration of separate tests to different groups of individuals and the results are "linked" to construct an item pool of the requisite size. Linking is necessary because the latent trait item parameter estimates are only invariant to within a linear transformation due to the arbitrary nature of the zero point and unit of measurement defined by the calibrations.

Despite the importance of linking to the construction of large item pools, minimal research has been done to determine how linking should be accomplished or how to determine the quality of the linking. Numerous methods have been proposed for linking tests (Warm, 1978; Lord, 1977; Marco, 1977; Wright, 1977; Rentz, 1978; Rentz and Bashaw, 1977), and the author has developed a few variations (a full list will be given later). However, these methods have not been thoroughly evaluated, and their controlling parameters are unknown. In many cases, when item pools are obviously constructed through a linking procedure, the actual procedure used is not mentioned as if it were not worthy of discussion (Valgeirslottir and Farr, 1978; Bejar, Weiss, and Kingsbury, 1977; Forster and Doherty, 1978; Forster, Ascher, and Karr, 1978). The research that does exist does not cover the area systematically, which leaves a disturbing gap in the methodology necessary to apply the latent trait models. It is the purpose of this paper to bridge the gap with a research structure that can support further applications of latent trait models. In analyzing the linking problem and reviewing the relevant research literature, four questions were identified that needed answers before linking can be performed with confidence.

Paper presented at the meeting of the American Educational Research Association, San Francisco, April, 1979. This research was supported by Contract Number N00014-77-C0097 from the Personnel and Training Research Programs of the Office of Naval Research.

They are as follows: (a) what sample size is required to adequately link tests? (b) how many test items should overlap between linked tests to allow accurate translation of scale? (c) are the existing item parameter estimates reduced in quality by the linking of new calibrations? and (d) what linking procedure is best for each of the two latent trait models being considered? These four questions will be addressed, based on the detailed analysis of a realistic application of several linking procedures.

Latent Trait Models

Two latent trait models have commonly been used in applications requiring large item banks (e.g., tailored testing, computer assisted test construction). They are the one- and three-parameter logistic models. Although some work has been done with normal ogive models, the research reported here will deal exclusively with these two logistic models.

The one-parameter logistic model is given by the equation

$$P(u_{ij}) = \frac{e^{u_{ij}(b_i - \theta_j)}}{1 + e^{(b_i - \theta_j)}} \qquad (1)$$

where u_{ij} is the score of Person j on Item i (either 0 or 1), $P(u_{ij})$ is the probability of the item score, b_i is the difficulty parameter for Item i, and θ_j is the ability parameter for Person j. This model assumes a unidimensional latent trait, local independence, no guessing, and equal discrimination for all items. Although these assumptions are not reasonable for many tests (i.e., multiple choice achievement tests), the model has been used with some success for a number of applications (Rentz and Bashaw, 1977; Ireland, 1976; Woodcock, 1972).

The three-parameter logistic model is given by the equation

$$P(u_{ij} = 1) = c_i + (1 - c_i) \frac{e^{Da_i(b_i - \theta_j)}}{1 + e^{Da_i(b_i - \theta_j)}} \qquad (2)$$

where u_{ij}, $P(u_{ij})$, b_i, and θ_j are as defined above, D is the constant 1.7 used to increase the similarity of the logistic function to the normal ogive function, c_i is the lower asymptote of the item characteristic curve for the item (pseudo-guessing level), and a_i is the discrimination parameter of the item. The model assumes that ability is unidimensional, and that local independence holds between item responses.

Linking Techniques

Four linking techniques were used for this study: (a) the major axis method, (b) the least squares method, (c) the least squares method with outliers deleted, and (d) the maximum likelihood method. The major axis method was the only one used with the one-parameter logistic model, while all four were used with the three-parameter logistic model. The reason only one technique was used with the one-parameter model was that good linking results had been obtained in the past using the major axis method,

obviating the need to check more complicated techniques. In contrast, an adequate procedure for the three-parameter model had not yet been determined.

Major axis method. The major axis method derives its name from the linear equations used to transform the parameter scales for the items from the various tests. These equations correspond to the equations for the major axis of the ellipse of the scatter plots made from the parameter estimates from two administrations of a set of items. This method gives the same result as the least squares method when the correlation between two sets of estimates of the item parameters is 1.0.

When the one-parameter logistic model difficulty parameters are being linked, the transformation is determined by the following procedures. First, the mean difficulty parameter estimate, $\bar{b}_B$, for the linking base test (the test that is used to define the parameter scale) is computed for the items in common between the base test and the test (Test 1) to be linked. Next, the mean difficulty parameter estimate for Test 1, $\bar{b}_1$, is computed. The linking constant used to determine the linking transformation is then computed by

$$k = \bar{b}_B - \bar{b}_1 \tag{3}$$

This constant is then added to all of the difficulty parameter estimates for the items on Test 1 to give transformed parameter estimates as shown by Equation 4.

$$b_{Ti} = b_i + k \tag{4}$$

The transformed values are then combined with the base test values for the common items using a weighted averaging procedure.

$$b'_{Bi} = \frac{n_B b_{Bi} + n_1 b_{Ti}}{n_B + n_1} \tag{5}$$

This last step was performed in an attempt to increase the precision of the parameter estimates as more test results were obtained. It should be noted that after Equation 5 is applied, parameter estimates from the items in common to the two tests are based on $n_B + n_1$ cases, while the rest of the items are based on only n_B or n_1 cases.

When the three-parameter logistic model parameters are linked using the major axis method, the procedure is more complicated because more parameters must be transformed and fewer assumptions are made concerning the relation between parameter estimates from different tests. The first step in the linking of the three-parameter model is to compute the mean and standard deviation of the difficulty parameters of the overlapping items in the linking base test, b_B and S_{bB}, and the test to be linked, b_1 and S_{b1}. To transform the parameters from Test 1 to the base test scale, the following equations are used:

$$b_{Ti} = \frac{S_{bB}}{S_{b1}} b_i - \bar{b}_1 \frac{S_{bB}}{S_{b1}} + \bar{b}_B \tag{6}$$

and

$$a_{Ti} = \frac{S_{bl}}{S_{bB}} a_i, \tag{7}$$

for the difficulty and discrimination parameters respectively. Note that the c-parameter scale does not have to be transformed since all of the parameters are already on the same 0-1 probability scale.

After the transformed values have been obtained, they are combined whenever previous estimates exist (i.e., for the overlapping items) using the analog to Equation 5. The same equation is used, but the appropriate a, b, and c parameters are substituted for the b-values from the one-parameter model.

Least squares method. Whereas the major axis method was derived algebraically assuming a correlation of 1.0 between the separate estimates of the item parameters, the least squares method assumes the correlation is less than 1.0, and that a regression procedure will minimize the error of estimates from the transformation. That is, the major axis method assumes the parameters are known, but are on a different scale, while the least squares method assumes the parameter estimates contain error that needs to be controlled.

The least squares method begins by computing the slope and intercept terms of the regression equation predicting the base test b-parameter values using b-values from the test to be linked, Test 1. The b-parameter estimates for the items in common to the two tests are used for this purpose. The Test 1 b-parameter estimates are then transformed to the base test scale using the regression equation.

$$b_{Ti} = \beta_o + \beta_1 b_i \tag{8}$$

Estimates for the common items are then combined using the weighted average procedure given in Equation 5.

To transform the a-parameters of the three-parameter logistic model, the means of the a-values for the items common to the two tests are computed, $\bar{a}_B$ and $\bar{a}_1$. A scaling constant is then computed by taking the ratio of these two values.

$$C = \frac{\bar{a}_B}{\bar{a}_1} \tag{9}$$

The transformed a-values are computed by multiplying the Test 1 values by the constant

$$a_{Ti} = Ca_i. \tag{10}$$

As with the b-values, the a-value estimates for the common items are combined using the weighted average procedure.

The c-parameter estimates are handled exactly the same way in the least squares method as in the major axis method.

Least Squares Outlier Deletion Method. The least squares outlier deletion method is the same as the least squares method with the exception that any common items with b-value estimates more than two standard-error-of-estimates away from the regression line are deleted from the computation

of the linking equations. After these items are dropped a new regression equation is computed for use in the actual linking process. The dropped items are not deleted from the item pool, however, they are merely dropped from the set of common items. They are considered as non-common items and their parameter values are transformed in the same manner as the other non-common items.

The outlier deletion method was developed because in small sample calibrations, the b-values obtained from the three-parameter model were occasionally well outside the range of what was expected (e.g., -32.076). It was hoped that by deleting these poorly estimated parameters from consideration in computing the linking equations, the overall quality of the linking would be improved.

Maximum Likelihood Method. The maximum-likelihood method can only be performed using the LOGIST program developed by Wood, Wingersky, and Lord (1976). In order to perform the linking using this program, the test data must first be edited into a large matrix that has the item numbers as the column indices, and examinees as the row indices. A representation of this matrix is presented in Figure 1. The editing program stores the responses from an examinee's response string in the appropriate locations of the matrix. For example, in Figure 1, Examinee 1 responded to Items 1, 3, and 4 out of the overall pool, while Items 2, 5 and 6 were not on the test. These latter items are coded as "not reached" in the data matrix. Examinee 100 took a different test and has responses from Items 3, 4, 5, and 6, but Items 1 and 2 were not included in that test and are therefore coded as "not reached". Once all of the data from the tests to be linked is edited into this matrix, the LOGIST program is run, yielding maximum likelihood estimates for the parameters on all of the test items. The fact that items are in common between pairs of tests insures that the parameter estimates are all on the same scale. The parameter estimates obtained are those that maximize the probability of the response data that has been input.

Insert Figure 1 about here

An important distinction is made in the LOGIST program between items that are "not reached" and those actively omitted. "Not reached" items are not included in the analysis of the data at all, while omitted items are assumed to be items that examinees do not know the answer to but could respond to at the chance level. Omitted responses are used in the estimation of the parameters.

Research Design

The linking procedures can be evaluated in at least two ways: through the use of simulation data, or through the use of live testing data. Using simulation data provides the advantage of being able to compare parameter estimates to the known item parameters used to generate the data. However, simulated data seldom represent live testing data adequately, particularly in factor structure and examinee variables such as guessing.

Evaluating the linking procedure using live testing data, on the other hand, gives a realistic check on procedures, but a good criterion for accuracy of parameter estimation is lacking since the true parameters are unknown. Instead, estimates are checked against large sample results, or the consistency of several analyses is used for evaluative purposes.

Obviously, neither empirical nor simulation procedures constitute the definitive evaluation technique. The best procedure would probably be to use both to evaluate the linking methods. However, due to time and effort constraints, only live testing data was used in the research reported here. This mode of evaluation was selected to give more directly applicable results than could be obtained from simulations.

The basic design used for the evaluation of the linking procedures was to sample short tests from a longer test in such a way that each short test selected had a predetermined number of items in common with the test previously sampled. Each short test was calibrated and linked using each of the linking methods mentioned above. The resulting parameter estimates were then compared to those obtained from a full calibration of the entire test using a large sample, the large test serving as the criterion by which to evaluate the linking.

Calibration Programs. Three calibration programs were used in conjunction with the linking methods: (a) the one-parameter logistic (Rasch), maximum likelihood program developed by Wright and Panchapakesan (1967) and modified by Reckase; (b) the three-parameter logistic, quasi-maximum likelihood program (LOGIST) developed by Wood, Wingersky and Lord (1976), and (c) the three-parameter logistic, ancillary estimation procedure (ANCILLES) developed by Croll & Urry (1978). The Wright and Panchapakesan program was selected since good results had been obtained in the past (Reckase, 1977). Two three-parameter logistic programs were used in the hope that one would require smaller samples than the other for linking purposes. The LOGIST program had been used with success in the past, but large samples were required for stable estimation.

Data. In order for the research design of this study to be successfully implemented, a long test was needed that had been administered to a relatively large number of examinees. Also, since the application of the linking models to achievement testing was of interest, a test that was representative of the factor structure of a typical achievement test was desired. The only data available to the author that fulfilled these criteria were the Iowa Tests of Educational Development, ITED (1972). This test is a general achievement measure covering seven subareas: expression, quantitative thinking, social studies, natural sciences, literature, vocabulary, and sources of information. Despite the fact that this test has seven distinct subtests, a composite score is produced which correlates highly with general measures of achievement, and it has a very high internal consistency reliability in most situations (e.g.,.96 for Iowa 12th graders). A principal component factor analysis of the test confirmed its multidimensional nature, but also showed that it had a strong first factor. Thus, this battery met the requirements for this study.

The ITED had 357 items, a length that made it ideal for this study, and a sample of 4,000 examinees was available. The sample was made up of 1,000 Iowa school students from each of grades 9, 10, 11, and 12 that took the test during the 1975-1976 school year. The distribution of total scores on the test had a strong negative skew with $\bar{X} = 184.61$ and SD = 61.51. The data were obtained through Dr. William Coffman, the director of the Iowa Testing Programs.

Procedure. In evaluating the linking methods, four variables were of interest. They were the linking methods themselves, the calibration procedures, the sample sizes required, and the number of items in common between two tests. In order to evaluate the effects of these variables,

data-sets of the sample sizes required and tests with the desired number of common items were produced. The data were than analyzed using the calibration and linking procedures.

The procedure for developing the data-sets was as follows:

1. Fifty items were selected using a stratified random sampling plan from the 355 items available for analysis (two items were discarded because the LOGIST program would not converge on parameter estimates). The strata used were the subtests of the ITED. This set of items was designated Test A.
2. From Test A, n items were randomly selected as common items. In this study, n was either 5, 15, or 25. Next, (50 - n) items were selected using a stratified random sample from the items not used in Test A. The total of fifty items, including the n common items, was designated Test B.
3. A set of n common items were then randomly selected from Test B. As above, (50 - n) items were then selected from the items not in Test B. Membership in Test A was ignored when making these selections. The 50 items selected were labeled Test C.
4. The same procedure was used to randomly select Test D.
5. Once the four tests with the three levels of overlap were selected, subsamples of the total 4,000 population were obtained using a systematic sampling procedure. That is, every nth case was taken so that samples of 100, 300, 500, 1000, and 2000 were obtained. Different first cases were used so that different examinees would be included on each of the four tests within a sample. The 2000 samples resulted in substantial overlap, but the analyses were still run for that case. Because of the non-independence of the samples, the results for the 2000 sample should be interpreted cautiously.
6. Each test was calibrated using each calibration program, with data of each sample size.
7. Each set of test calibrations was linked for each sample size and overlap using each linking procedure.

Analysis. The linked parameter estimates obtained from each of the combinations of sample size, overlap, and method were analyzed for quality in three ways. First, the linked parameter estimates were correlated with the estimates obtained from the full 357 item test using the full sample of 4,000. These correlations were then compared using a procedure given in Snedecor and Cochran (1967) for determining whether n correlations come from the same population. The usual Fisher r to z transformation was used to test whether specific pairs of correlations were significantly different.

The second type of analysis performed was the computation of the sum-of-squared deviations quality of linking index suggested by Wright (1977) for use with the one-parameter logistic model. This index is based on the formula

$$S^2_{BT} = \Sigma (b_{Bi} - b_{Ti})^2 \qquad (11)$$

where b_{Bi} and b_{Ti} are defined above. The summation is over the common items in the base test and the test to be linked. This statistic can be transformed to a χ^2 statistic with K-1 degrees of freedom using the

formula

$$\chi^2 = \frac{S_{BT}^2 NK}{12} \qquad (12)$$

where K is the number of common items and N is the sample size for Test 1. This statistic was computed for both the one-parameter and three-parameter b-values, although its applicability to the three-parameter model has yet to be determined.

Finally, scatter plots were constructed comparing the large sample parameter estimates to the linked parameter estimates as a check for any non-linearity. If a curvilinear relationship were suspected, tests for deviations from linearity were performed to determine the seriousness of the deviations.

Results

The results section is organized as follows. First, the correlations and sum-of-squared deviations data for the one-parameter linking and their analysis will be presented. Next, the results of the three-parameter linkings will be given. Finally, the results of the two models will be compared where appropriate. The correlations between the linked one-parameter logistic difficulty parameter sets obtained using the major axis method and the difficulty parameter estimates obtained from the calibration of the full ITED are presented in Table 1. The correlations are shown for the linkings using 5, 15, and 25 overlapping items, and for sample sizes of 100, 300, 500, 1000, and 2000. In addition, the correlations are presented for the linking results after two tests have been combined, then after a third test has been added, and finally after a fourth test has been added. The data were analyzed in this way to check for the drift mentioned by Rentz (1978).

Insert Table 1 about here

An initial impression obtained from looking at Table 1 was that all of the correlations were high indicating that the linking was working well. The scatter plots corresponding to these correlations (not shown) also showed that the plots were linear with slope 1.0 indicating that only an additive constant was needed to transform the parameter estimates to those obtained from the full test calibration, as would be predicted by the model. A check for significant differences among the sets of correlations using the procedure outlined in Snedecor and Cochran (1967, pp. 186-187) also gave predictable results. For all three levels of item overlap, the correlations increased significantly with increased sample size. The biggest differences were found between the 100 and 300 samples. Significant differences were also found across levels of overlap at the 100, 1000, and 2000 sample sizes.

The results showed that there was a tendency for the correlations to increase with increased overlap, although the differences may have no practical meaning despite their statistical significance since the correlations were so close to 1.0. No significant change in the magnitude of the correlations was found as the number of tests linked together went from two to four.

The results for the sum-of-squared deviations quality of linking index are presented in Table 2 for the same treatment combinations as the correlations. The values presented in the table are the χ^2 values corresponding to the squared deviations. In all but three cases, these χ^2 values were definitely significant. The significance was very high even for the conditions where the correlations were .999. In general, this statistic seems to have little relationship to the quality of the linking, so the values obtained for the three-parameter model analyses will not be reported. In any event, the application of the statistic to those parameters was questionable.

Insert Table 2 about here

The linking results for the major axis method when applied to the three-parameter logistic model are presented in Table 3. The correlational results for both the ANCILLES and LOGIST calibrated items are presented in the same table so a direct comparison can be made. Note that the correlations were computed separately for each of the three-parameters of the logistic models (see Equation 2) and for each sample size. Only the 15 item overlap data is presented for the three-parameter models because complete data for the other overlaps could not be obtained in time for this paper. Results of the 5 and 25 item overlaps will be published in a technical report later in the year.

Insert Table 3 about here

The same types of correlational comparisons were performed for the three-parameter linkings as were done for the one-parameter linkings. First, a check was made to determine if the correlations with the full calibration changed as more tests were linked together. Only one significant change was noted among the AB, ABC, and ABCD correlations, and that was for LOGIST a-values for the 300 sample ($\chi^2_{(2)} = 23.11$). Since this was the only significant difference out of the 30 tests it was probably a chance value. Thus there does not seem to be any "drift" in parameter estimates as more items were added to those previously linked.

The second tests performed were to determine whether quality of the linking improved with increased sample size. As with the one-parameter model, the correlations showed a substantial improvement with increased sample as shown by the χ^2 values at the bottom of the table. The LOGIST based linkings, however, did not improve as consistently as those based on the ANCILLES calibrations. The final comparison made on the results of the major axis linking was to determine which set of calibrated results more closely matched the full scale calibration. The underlined values in the ABCD columns indicate which of the corresponding correlations were significantly greater in a pair-wise comparison. Although the multiple comparisons used poorly controlled the error rate of the significance tests, the results seem to clearly show that the major axis method applied to the ANCILLES calibrations yield better results than for the LOGIST calibrations. Note, however, that the relationship with the total test linking for the major axis ANCILLES results was best for the b-parameters with substantially lower correlations for the a- and c-parameters--especially

at the smaller sample size.

Table 4 presents the results of the least squares linking method for the two calibration procedures using the same format as Table 3. As for the previous method, no significant change was found in the correlations as additional tests were linked to the previously linked set. Also, as before, a significant increase was found in the correlations with increased sample size. The LOGIST correlations again did not improve as consistently as the ANCILLES correlations.

Insert Table 4 about here

When the overall quality of the linkings was compared by testing each pair of correlations in the ABCD sets, the results were not as clear cut as for the major axis method. The ANCILLES based procedure did better for eight of the parameter linkings (as shown by the underlinings) and the LOGIST based procedure did better for six. In one case there was no significant difference. Due to the high error rate for multiple comparisons, neither procedure can be said to be superior. It is interesting to note, however, that the LOGIST based procedure did better for the a-parameters, while the ANCILLES based procedure did better for the b- and c-parameters.

The results of the evaluation of the linking for the final two procedures are presented in Table 5. The correlations with the full test calibrations are presented for the least-squares outlier deletion procedure and the maximum likelihood procedure. In both cases the LOGIST program was used to obtain item parameter estimates. Since the maximum likelihood procedure calibrates all of the tests to be linked in one step, only one column of correlations was presented. The outlier deletion procedure did add tests one at a time, and no significant change in the correlations was found with the added tests. Thus no evidence of drift in calibrations was found for any of the linking techniques evaluated.

Insert Table 5 about here

Both the outlier deletion and maximum likelihood methods resulted in an increase in quality as the sample size increased, as indicated by the χ^2 values at the bottom of Table 5. This was another finding common to all of the procedures. The comparison of the overall linkings between these two methods showed that the outlier deletion method was superior in nine cases, while the maximum likelihood procedure was better in six. Again, the high error rate of the multiple significance tests used makes it impossible to say which procedure is better. The maximum likelihood procedure was better for the 100 sample size and for b-parameter estimates, while the outlier deletion method was better for a- and c-parameter estimates.

Insert Table 6 about here

The obvious question now arises as to which is the best of the six three-parameter logistic linking procedures presented. Comparisons were run among the ABCD linking correlations to determine which procedure had the strongest relationship to the full test calibrations. The results showed no clearly superior method. The major axis linking of the ANCILLES calibration had the highest correlations for the b- and c-parameters, while the least squares linking of the LOGIST calibrations had the highest

correlations for the a-parameters. While these two procedures tended to perform better than the rest, only the major axis method applied to the LOGIST calibrations was clearly inferior.

The final comparison that was made in this study was between the one-parameter logistic and three-parameter logistic linking results. Significance tests were performed between the correlations for the linking of the b-parameters for the one-parameter model and the major axis linking of the ANCILLES calibrations. In all cases, the one-parameter correlations were significantly larger than the ANCILLES correlations. In fact, a sample of 500 was required for the ANCILLES linking before the correlation exceeded that for the 100 one-parameter logistic linking. This shows the obvious sample size advantage of the more simple logistic model.

Discussion

The purpose of this paper was to analyze six three-parameter logistic linking techniques and one one-parameter logistic linking technique to determine (a) the sample size required to adequately link tests, (b) the number of overlapping items needed to link tests, (c) whether parameter estimate drift occurs when new tests are linked to an existing pool, and (d) which procedure was best for linking item parameter estimates obtained from the three-parameter logistic model. In regard to the first of these points, the sample size required for linking, all procedures improved significantly with increased sample size. However, the one-parameter model improved most between the 100 and 300 samples and leveled out afterwards suggesting that a sample of 300 or more should be used for a stable linking. For the three-parameter linkings, the greatest improvement also occurred between the 100 and 300 samples, however the quality of the linking was still not very high for the 300 sample. A sample size of 1000 was required before the three-parameter linkings had any semblance of stability. If one were forced to link small sample size calibrations based on the three-parameter logistic model, the maximum likelihood procedure would probably be recommended since it gave the best results at the 100 sample size. However, the maximum likelihood method tended to yield extreme b-values and a-values of .01 for some items when small samples were used. These items should be dropped and the tests recalibrated to obtain a more stable result.

The results pertinent to the number of overlapping items required for a valid linking can only be discussed relative to the one-parameter model. Although an overlap effect was present for some sample sizes, it only seemed to be important at the 100 sample size. At that sample size, the 25 item overlap yielded a slightly better linking than the 5 and 15 item overlaps, suggesting that if small samples must be used, there should be more items in common between the tests to be linked. When large sample sizes were used, the effect of varying the number of overlapping items was minimal, although sometimes significant. The difference between correlations of .994 and .998 does not seem to be important, however. These results indicate that if ample samples of examinees are used for calibration, the tests to be linked need relatively few items in common.

The third area of interest in this research was to determine whether the quality of the linked parameter estimates improved or deteriorated as additional tests were linked into an item pool. The results obtained show that there is no change in quality as new tests are linked to previous

results. These findings are quite different from those obtained by Rentz (1978), but his tests were obtained over a period of years instead of from a single large test administration. The results of this study suggest that parameter drift is not due to an accumulation of error as the techniques are applied successively, but probably is caused by a change in the meaning of the items. In effect, the dimensionality of some items may be changing over time as the knowledge students bring to the testing session changes. Persons constructing items over a long period of time should check for possible "drift" effects.

The final, and probably the most important, question addressed by this paper is which of the various linking techniques is best for the three-parameter logistic model. Unfortunately, a clear answer cannot be given to this question. Initially, the major axis technique applied to the ANCILLES calibrations seemed to be the recommended procedure, but a good proportion of the advantage of the procedure is based on the c-parameter estimate correlations. A problem with correlational comparisons is that they require variations in the parameters for the comparisons to be meaningful. The LOGIST calibration procedure puts severe restraints on the variation in the c-parameter estimates making high correlations impossible even though the estimates are very similar. The choice of a procedure then depends on whether accuracy in b-parameter estimates is more or less important than accuracy in a-parameter estimates. If the former is the case the major axis ANCILLES procedure should be selected; in the latter case, the least squares LOGIST procedure is preferred. Unfortunately, both the a- and b-parameters are important when selecting items to maximize information. At the recommended sample size of 1000 the maximum likelihood approach seems to be a good compromise between the methods, giving fairly good, but not the best, correlations for [illegible] parameter. Maximum likelihood is also the most convenient technique to apply since the linking is performed in a single step.

In summary, one one-parameter logistic and six three-parameter logistic linking procedures have been evaluated on their ability to reproduce a large sample calibration of 357 items with calibrations using 50 items various combinations of sample size and overlap. The one-parameter linking did well regardless of the number of items in common between tests if sample sizes of 300 or more were used. The three-parameter linkings required a substantially larger sample size for stable linkings. A sample of 1000 was recommended. No clearly superior three-parameter linking procedure was found, although the major axis method applied to LOGIST calibrations was less accurate than the others. At the recommended 1000 sample size, the major axis ANCILLES method and maximum-likelihood LOGIST method seem to give a good combination of accuracy for estimation of the three-parameters. All of these conclusions were determined based upon multivariate achievement test data. If more unidimensional data were used, larger correlations would be expected.

References

Bejar, I. I., Weiss, D. J. and Kingsbury, G. G. Calibration of an item pool for the adaptive measurement of achievement (Research Report 77-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, September, 1977.

Croll, P. R. and Urry, V. W. ANCILLES: A program for estimation of the item parameters of normal ogive and logistic mental test models. Washington, D.C.: Research Section, Personnel Research and Development Center, U.S. Civil Service Commission, June, 1978.

Forster, F., Ascher, G., and Karr, C. Research on the Rasch measurement model. Paper presented at the annual meeting of the American Educational Research Association, Toronto, March, 1978.

Forster, F. and Doherty, V. Using the Rasch approach to measurement to solve practical school testing problems. Paper presented at the annual meeting of the American Educational Research Association, Toronto, March, 1978.

Ireland, C. M. An application of the Rasch one-parameter logistic model to individual intelligence testing in a tailored testing environment. Unpublished doctoral dissertation. Univeristy of Missouri-Columbia, 1976.

Jensema, C. J. Bayesian tailored testing and the influence of item bank characteristics. Proceedings of the first conference on computerized adaptive testing. United States Civil Service Commission, Washington, 1975.

Lord, F. M. Applications of item response theory to practical testing problems. Unpublished manuscript, 1977.

Marco, G. L. Item characteristic curve solutions to three interactable testing problems. Journal of Educational Measurement, 1977, 14(2), 139-160.

Reckase, M. D. The effect of item pool characteristics on the operation of a tailored testing procedure. Paper presented at the Spring meeting of the Psychometric Society, Murray Hill, N. J., June, 1976.

Reckase, M. D. Ability estimation and item calibration using the one- and three-parameter logistic models: a comparative study. (Research Report 77-1). Columbia, Mo.: University of Missouri, November, 1977 (NTIS No. AD A047943).

Rentz, R. R. Monitoring the quality of an item-pool calibrated by the Rasch model. Paper presented at the Annual meeting of the National Council on Measurement in Education, Toronto, March, 1978.

Rentz, R. R. and Bashaw, W. L. The national reference scale for reading: an application of the Rasch model. Journal of Educational Measurement, 1977, 14(2), 161-180.

Snedecor, G. W. and Cochran, W. G. Statistical methods. Ames, Iowa: The Iowa State University Press, 1967.

Valgeirslottir, S. and Farr, S. D. Measuring abilities of Icelandic youth. Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, March, 1978.

Warm, T. A. A primer of item response theory (Technical Report 941078). Oklahoma City, Ok.: U.S. Coast Guard Institute, November, 1978.

Wood, R. L., Wingersky, M. S. and Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (Research Memorandum RM-76-6). Princeton, NJ: Educational Testing Service, June, 1976.

Woodcock, R. W. Woodcock reading mastery tests. Circle Pines, Minn.: American Guidance Service, 1972.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14(2), 97-116.

Wright, B. D. and Panchapakesan, N. A procedure for sample free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

Table 1
Correlations of Linked One-Parameter Difficulty Estimates with Full Test Difficulty Estimates by Sample Size and Overlap
Major Axis Method

Sample Size	Overlap/Tests Linked: 5 AB	5 ABC	5 ABCD	15 AB	15 ABC	15 ABCD	25 AB	25 ABC	25 ABCD	χ^2	df
100	.949	.950	.958	.950	.948	.943	.972	.965	.969	6.54*	2
300	.980	.981	.984	.981	.982	.983	.986	.985	.988	2.13	2
500	.979	.982	.988	.991	.991	.992	.992	.992	.993	5.67	2
1000	.989	.991	.994	.997	.997	.997	.997	.997	.998	21.42***	2
2000	.998	.998	.998	.998	.998	.999	.999	.999	.999	11.89***	2
Items correlated	99	134	165	81	116	144	75	98	115		
χ^2			211.71***			345.93**			219.70***		
df			4			4			4		

* p < .05
** p < .01
*** p < .005

Table 2
Quality of Linking Statistic for the
One-Parameter Linking by
Sample Size and Overlap

Sample Size	Overlap/Tests Linked: 5 AB	5 ABC	5 ABCD	15 AB	15 ABC	15 ABCD	75 AB	75 ABC	75 ABCD
100	13.0	170.5	404.7	345.3	143.4	354.4	381.2	401.3	867.1
300	8.3*	143.2	214.7	115.9	276.3	303.1	410.6	536.5	710.0
500	18.5	182.0	460.9	251.6	150.4	362.3	587.4	723.1	797.8
1000	17.5	111.2	597.1	147.7	143.8	271.3	564.7	554.7	811.6
2000	18.8	13.7*	305.3	.2*	132.4	160.2	481.1	373.0	492.7
Overlapping items	5	11	19	15	19	22	25	27	33
$\chi^2(.05)$	9.49	18.31	28.87	23.68	28.87	33.92	36.42	40.11	55.76

*These values were not significant at the .05 level. All others indicated a significant difference from a quality linking..

Table 3
Correlations of Linked Three-Parameter Estimates
with Full Test Parameter Estimates
for the Major Axis Method 15 Item Overlap

Sample Size	Parameter	Calibration Procedure: ANCILLES AB	ANCILLES ABC	ANCILLES ABCD	LOGIST AB	LOGIST ABC	LOGIST ABCD
100	n_i	82	113	141	85	116	144
	a	.314	.330	.421	.600	.414	.367
	b	.766	.788	.772	.762	.733	.734
	c	.493	.484	.452	-.132	.004	-.017
300	a	.698	.687	.683	.827	.511	.533
	b	.937	.924	.930	.909	.900	.907
	c	.710	.600	.566	.549	.375	.344
500	a	.664	.718	.756	.523	.520	.439
	b	.965	.959	.963	.824	.809	.829
	c	.717	.689	.742	.381	.486	.422
1000	a	.830	.870	.888	.632	.679	.630
	b	.965	.968	.969	.808	.826	.842
	c	.743	.734	.769	.329	.539	.563
2000	a	.916	.926	.923	.934	.948	.945
	b	.969	.977	.978	.987	.988	.989
	c	.769	.805	.815	.531	.676	.588
$\chi^2_{(4)}$	a			119.29***			182.04***
	b			129.85***			240.91***
	c			42.01***			43.80***

***p < .005

Note: The row labeled n_i indicates the number of items used to compute the correlations. The ANCILLES program has fewer values because it automatically deletes poor items. The underlined values indicate cases where the one procedure was significantly better than the other.

Table 4
Correlations of Linked Three-Parameter Estimates
with Full Test Parameter Estimates
for the Least Squares Method 15 Item Overlap

Calibration Procedure

Sample Size	Parameter	ANCILLES AB	ANCILLES ABC	ANCILLES ABCD	LOGIST AB	LOGIST ABC	LOGIST ABCD
	n_i	82	113	141	85	116	144
	a	.394	.369	.419	.589	.695	.713
100	b	.709	.672	.670	.707	.670	.662
	c	.493	.484	.452	-.132	.004	-.017
	a	.676	.682	.676	.859	.850	.837
300	b	.950	.932	.928	.909	.888	.856
	c	.710	.600	.566	.550	.375	.344
	a	.612	.687	.726	.777	.835	.819
500	b	.960	.951	.952	.808	.789	.801
	c	.717	.689	.742	.381	.486	.422
	a	.864	.877	.892	.921	.925	.922
1000	b	.961	.964	.965	.796	.796	.806
	c	.743	.734	.769	.329	.539	.563
	a	.916	.926	.924	.947	.957	.954
2000	b	.970	.977	.978	.987	.988	.989
	c	.769	.805	.815	.531	.676	.588
$\chi^2_{(4)}$	a			123.00***			85.49***
	b			167.79***			278.44***
	c			41.15***			43.80***

***p < .005

Note: The row labeled n_i indicates the number of items used to compute the correlations. The ANCILLES program has fewer values because it automatically deletes poor items. The underlined values indicate cases where the ons procedure was significantly better than the other.

Table 5
Correlations of Linked Three-Parameter Estimates
with Full Test Parameter Estimates
for the Outlier Deletion and Maximum Likelihood Methods
15 Item Overlap

LOGIST

Sample Size	Parameter	Outlier Deletion AB	Outlier Deletion ABC	Outlier Deletion ABCD	Maximum Likelihood
	n_i	85	116	144	144
	a	.585	.692	.711	.735
100	b	.693	.633	.633	.673
	c	-.132	.004	-.017	.096
	a	.858	.850	.836	.793
300	b	.908	.889	.895	.943
	c	.550	.375	.344	.262
	a	.778	.835	.818	.811
500	b	.809	.786	.773	.923
	c	.381	.486	.422	.269
	a	.921	.925	.923	.917
1000	b	.796	.810	.800	.951
	c	.329	.539	.563	.400
	a	.947	.960	.954	.945
2000	b	.987	.988	.989	.983
	c	.531	.676	.588	.407
$\chi^2_{(4)}$	a			86.00***	73.16***
	b			297.13***	180.97***
	c			43.80***	10.65*

*p < .05
***p < .005
Note: The row labeled n_i indicates the number of items used to compute the correlations. The underlined values indicate cases where one procedure was significantly better than the other.

Table 6
Correlations with the Full Test Calibrations
for Five Three-Parameter Logistic Linking Procedures

Sample Size	Parameter	Calibration Procedure: LOGIST — Outlier Deletion	LOGIST — Maximum Likelihood	LOGIST — Least Squares	ANCILLES — Major Axis	ANCILLES — Least Squares
100	a	.711	.735	.713	.421	.419
	b	.633	.673	.662	.772	.670
	c	-.017	.096	-.017	.452	.452
300	a	.836	.793	.837	.683	.676
	b	.895	.943	.896	.930	.928
	c	.344	.262	.344	.566	.566
500	a	.818	.811	.819	.756	.726
	b	.773	.923	.801	.963	.952
	c	.422	.269	.422	.742	.742
1000	a	.923	.917	.922	.888	.892
	b	.800	.951	.806	.969	.965
	c	.563	.400	.563	.769	.769
2000	a	.954	.945	.954	.923	.924
	b	.989	.983	.989	.978	.978
	c	.588	.407	.588	.815	.815

Note: The largest correlation in each row is underlined.

Figure 1
Data matrix for
Linking with the LOGIST PROGRAM

Examinee No.	Item No. 1	2	3	4	5	6 . . .	K
1	1	NR	0	0	NR	NR	
2	1	NR	1	1	NR	NR	
3							
4							
.							
.							
.							
100	NR	NR	0	0	1	1	
101	NR	NR	1	0	0	1	
.							
.							
.							
N							